Test-retest reliability of structural brain networks from diffusion MRI

Test-retest reliability of structural brain networks from diffusion MRI

Colin R. Buchanan[a,b], Cyril R. Pernet[c,d], Krzysztof J. Gorgolewski[e], Amos J. Storkey[b], Mark E. Bastin[c,d,*]

[a] *Doctoral Training Centre in Neuroinformatics and Computational Neuroscience, School of Informatics, University of Edinburgh, Edinburgh, UK*
[b] *Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, Edinburgh, UK*
[c] *Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK*
[d] *Brain Research Imaging Centre, Neuroimaging Sciences, University of Edinburgh, Edinburgh, UK*
[e] *Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*

Abstract

Structural brain networks constructed from diffusion MRI (dMRI) and tractography have been demonstrated in healthy volunteers and more recently in various disorders affecting brain connectivity. However, few studies have addressed the reproducibility of the resulting networks. We measured the test-retest properties of such networks by varying several factors affecting network construction using ten healthy volunteers who underwent a dMRI protocol at 1.5 T on two separate occasions.

Each T_1-weighted brain was parcellated into 84 regions-of-interest and network connections were identified using dMRI and two alternative tractography algorithms, two alternative seeding strategies, a white matter waypoint constraint and three alternative network weightings. In each case, four common graph-theoretic measures were obtained. Network properties were assessed both node-wise and per network in terms of the intraclass correlation coefficient (ICC) and by comparing within- and between-subject differences.

Our findings suggest that test-retest performance was improved when: 1) seeding from white matter, rather than grey; and 2) using probabilistic tractography with a two-fibre model and sufficient streamlines, rather than deterministic tensor tractography. In terms of network weighting, a measure of streamline density produced better test-retest performance than tract-averaged diffusion anisotropy, although it remains unclear which is a more accurate representation of the underlying connectivity. For the best performing configuration, the global within-subject differences were between 3.2% and 11.9% with ICCs between 0.62 and 0.76. The mean nodal within-subject differences were between 5.2% and 24.2% with mean ICCs between 0.46 and 0.62. For 83.3% (70/84) of nodes, the within-subject differences were smaller than between-subject differences. Overall, these findings suggest that while current techniques produce networks capable of characterising the genuine between-subject differences in connectivity, future work must be undertaken to improve network reliability.

Keywords: connectome, diffusion MRI, human brain, network, test-retest, tractography

*Corresponding author: Dr Mark E. Bastin, Brain Research Imaging Centre, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK.

Email address: Mark.Bastin@ed.ac.uk (Mark E. Bastin)

1. Introduction

The structural connectome (Hagmann, 2005; Sporns et al., 2005) can be explored at a macroscopic scale *in vivo* through diffusion MRI (dMRI) and whole-brain tractography. Under this approach, segmented cortical regions, for example Brodmann areas, form the nodes of a network whilst tractography is used to construct a set of white matter fibre tracts which form the connections. Such techniques have the potential to map the collective wiring of many billions of axonal fibres and may provide detailed information on how cerebral white matter structure correlates with function, and potentially dysfunction, in both health and disease.

Although there is currently no accepted method for constructing dMRI structural networks, previous approaches have typically followed a similar organisation. Firstly, network nodes are formed from segmentation of high resolution 3D T_1-weighted volume scans, often by registration to neuroanatomical atlases (Maldjian et al., 2003; Shattuck et al., 2008; Tzourio-Mazoyer et al., 2002) or surface parcellation based on cortical sulci and gyri (Desikan et al., 2006; Fischl et al., 2004b). The number and choice of nodes requires careful consideration as this affects the resulting measures of connectivity (Zalesky et al., 2010b). Previous approaches have typically divided the cortex into fewer than 100 grey matter nodes, though some researchers have used finer parcellations with thousands of nodes of roughly uniform size, primarily to estimate global network properties (Cammoun et al., 2011; Hagmann et al., 2007, 2008; Zalesky et al., 2010b). Secondly, cross-modal registration (Andersson et al., 2007; Greve and Fischl, 2009; Jenkinson and Smith, 2001; Jenkinson et al., 2002) is typically required to align cortical labels to diffusion space. Thirdly, either deterministic (Basser et al., 2000; Lazar et al., 2003; Mori et al., 1999) or probabilistic (Behrens et al., 2003b, 2007; Parker et al., 2003) tractography is used to construct white matter tracts from dMRI data. Lastly, network connections are computed by quantifying tracts connecting between regions. Network weights typically reflect a count of interconnecting tracts (Hagmann et al., 2008) or some measure of tissue microstructure, such as diffusion anisotropy, averaged along the length of each tract (Iturria-Medina et al., 2007; Robinson et al., 2010). Connections may then be assessed directly in group-control studies (Zalesky et al., 2010a) or network measures derived from graph-theory (Rubinov and Sporns, 2010) may be used to characterise patterns of connectivity in individuals or across populations.

The first connectome dMRI studies demonstrated whole-brain network analysis in healthy volunteers (Hagmann et al., 2007, 2008). Various organisational properties have since been reported, such as the identification of highly connected 'hub' nodes, a modular structure and 'small-world' organisation (Hagmann et al., 2008; Honey et al., 2008; Sporns, 2011; Van Den Heuvel and Sporns, 2011; Yan et al., 2011). Recent studies have assessed structural connectivity in normal ageing (Gong et al., 2009; Robinson et al., 2010; Ystad et al., 2011; Wen et al., 2011), Alzheimers disease (Lo et al., 2010), mild cognitive impairment (Wee et al., 2011), stroke (Crofts et al., 2010), amyotrophic lateral sclerosis (Verstraete et al., 2011), multiple sclerosis (Shu et al., 2011) and neuropsychiatric disorders (Skudlarski et al., 2010; Zalesky et al., 2011). However, only a small subset of studies have assessed the reliability of the resulting networks (Bassett et al., 2010; Cammoun et al., 2011; Cheng et al., 2012; Hagmann et al., 2008; Vaessen et al., 2010; Zalesky et al., 2010b), and currently there is a lack of assessment concerning the reproducibility of these approaches.

In this paper, we constructed networks from repeat scans of healthy volunteers by varying several factors affecting the construction of networks. We compared two alternative tractography algorithms (deterministic and probabilistic), two seeding approaches (grey and white matter), and three alternative network weightings (streamline density, streamline density with length correction and a measure of tract-averaged diffusion anisotropy). We also investigated whether false connections could be reduced by an anatomically motivated filtering of streamlines based on length in white matter. In each case, we then quantified the reliability of four graph-theoretic measures using the intraclass correlation coefficient (ICC) and by comparing within- and between-subject average differences. Since these measures are an essential prerequisite for more complex analyses, such as small-world measures or the identification of network hubs, their reliability is crucial to the ultimate interpretation of such networks.

2. Materials and methods

An automated connectivity mapping pipeline was developed to construct white matter structural networks from dMRI data using Nipype ('Neuroimaging in python pipelines and interfaces': `http://nipy.sourceforge.net/nipype`; Gorgolewski et al. 2011), a framework which integrates a number of neuroimaging toolkits. The steps within this framework are detailed in the following sections.

2.1. Cohort

Ten healthy volunteers (six female) aged between 50 and 58 years underwent a dMRI protocol on two separate occasions over an interval of either two or three days (Gorgolewski et al. 2013; data available to download from GigaDB; `http://gigadb.org`). The study was approved by the local research ethics committee and informed consent was obtained from each subject.

2.2. MRI protocol

All imaging data were acquired using a GE Signa HDxt 1.5 T (General Electric, Milwaukee, WI, USA) clinical scanner with a manufacturer supplied 8-channel phased-array head coil. For the dMRI protocol, single-shot spin-echo echo-planar (EP) diffusion-weighted whole-brain volumes ($b = 1000\ s\ \mathrm{mm}^{-2}$) were acquired in 64 non-collinear directions, along with seven T_2-weighted volumes ($b = 0\ s\ \mathrm{mm}^{-2}$) (Jones et al., 2002). The repetition and echo times were 16.5 s and 98.3 ms respectively. Seventy-two contiguous axial 2 mm thick slices were acquired resulting in 2 mm isotropic voxels. In the same session, high resolution 3D T_1-weighted inversion-recovery prepared, fast spoiled gradient-echo volumes were acquired in the coronal plane with 180 contiguous 1.3 mm thick slices resulting in voxel dimensions of $1 \times 1 \times 1.3$ mm.

2.3. Neuroanatomical segmentation

In order to form a corresponding set of network nodes across subjects, each T_1-weighted brain was divided into distinct neuroanatomical regions. For this purpose, volumetric segmentation and cortical reconstruction was performed with the FreeSurfer image analysis suite (`http://surfer.nmr.mgh.harvard.edu`) using the default parameters. The Desikan-Killiany atlas delineated 34 cortical structures per hemisphere (Desikan et al., 2006; Fischl et al., 2004b). Additionally, sub-cortical segmentation was applied to obtain eight grey matter structures per hemisphere: accumbens area, amygdala, caudate, hippocampus, pallidum, putamen, thalamus and ventral diencapahlon (Fischl et al., 2002, 2004a). Following segmentation, 84 grey matter regions were retained per subject. The results of the segmentation procedure were used to construct grey and white matter masks for each subject.

2.4. Diffusion processing

Using tools freely available in the FSL toolkit (FMRIB, Oxford University: `http://www.fmrib.ox.ac.uk`), the dMRI data underwent eddy current correction to account for systematic imaging distortions and bulk patient motion using affine registration to the first T_2-weighted EP volume of each subject (Jenkinson and Smith, 2001). Diffusion tensors were fitted at each voxel location and fractional anisotropy (FA) values estimated (Basser and Pierpaoli, 1996). FA values express the degree of anisotropic diffusion per voxel and are considered to reflect the underlying fibre density. Skull stripping and brain extraction were performed on the T_2-weighted EP volumes acquired along with the dMRI data and applied to the FA volume of each session (Smith, 2002). All diffusion parameters and tractography were computed in native space.

2.5. Image registration

A cross-modal nonlinear registration protocol was employed to align neuroanatomical ROIs from T_1-weighted volume to diffusion space. Firstly, linear registration (FLIRT: Jenkinson and Smith 2001; Jenkinson et al. 2002) was used to initialise the alignment of each brain-extracted FA volume to the corresponding FreeSurfer extracted brain using a mutual information cost function and an affine transform with 12 degrees of freedom. Following this initialisation, a nonlinear deformation field based method (FNIRT: Andersson et al. 2007) was used to refine local alignment. FreeSurfer segmentations and anatomical labels were aligned

to diffusion space by applying these transforms using nearest neighbour interpolation. For visual inspection of the registration accuracy, each T_1-weighted extracted brain was also aligned to diffusion space using tri-linear resampling. FA volumes were chosen as the target, rather than T_2-weighted ($b = 0$) volumes, as visual inspection revealed greater registration accuracy using FA. For each subject, a binary brain mask was formed in diffusion space from all grey and white matter voxels obtained from FreeSurfer. These masks were used to constrain tractography to white matter structures and targeted cerebral voxels.

2.6. Tractography

We used two alternative tractography algorithms, one based on deterministic tensor tractography (FACT: Mori et al. 1999; Cook and Alexander 2006), and a probabilistic algorithm modelling two fibre directions at each voxel (FDT BedpostX/ProbtrackX: Behrens et al. 2003b, 2007). The deterministic approach estimates the best fit of the diffusion tensor model at each voxel, whereas the probabilistic approach estimates a distribution of possible orientations. For FACT, streamlines were constructed from voxel to voxel following the principal directions of diffusion estimated from the diffusion tensor. For the probabilistic approach, the distributions for tracking were generated with a two-fibre model per voxel (Behrens et al., 2007). Streamlines were then constructed by sampling from these distributions during tracking using 100 Markov Chain Monte Carlo (MCMC) iterations with a fixed step size of 0.5 mm between successive points. The termination criteria was the same for both algorithms, for which a streamline was terminated by exceeding a curvature threshold of 80 degrees, entering a voxel with FA below 0.1, or encountering an extra-cerebral voxel. The values of the curvature and anisotropy constraints were set empirically.

For each tractography algorithm we employed two alternative seeding approaches, namely, *white matter seeding* (WM-seeding) and *grey matter seeding* (GM-seeding). Under WM-seeding, tracking was initiated from all white matter voxels and streamlines were constructed in two collinear directions until terminated by the stopping criteria. Under the GM-seeding approach, tracking was initiated from all grey matter voxels (within an ROI) and streamlines were constructed in a single direction until terminated by the same stopping criteria.

2.7. Network construction

Connections between regions were computed by identifying the streamlines connecting each pair of grey matter ROIs. The endpoint of a streamline was considered to be the first grey matter ROI encountered when tracking from the seed location, subject to waypoint constraints. Streamlines which did not connect to an ROI were discarded. Networks were computed for 13 different thresholds of streamline filtering by minimum contiguous length in white matter, from 0 to 6.0 mm in increments of 0.5 mm. For instance, a threshold of l mm discards any streamline which does not pass through at least l mm in white matter between grey matter ROIs. The white matter regions obtained from FreeSurfer were used as the waypoint mask.

2.8. Network weighting

Three types of network weighting were recorded for each set of streamlines, two based on streamline density and a third on tract-averaged FA. In each case, connections were recorded in an $n \times n$ adjacency matrix, where the entry a_{ij} denotes the connection (edge) weight between node i and node j. The first weighting, termed *streamline density* (SD-weighted), records the interconnecting streamline density corrected for ROI size,

$$a_{ij} = \frac{2}{g_i + g_j}|S_{ij}|, \tag{1}$$

where S_{ij} is the set of all streamlines found between node i and node j (and $S_{ij} = S_{ji}$), and g_i and g_j are the number of grey matter voxels in nodes i and j. The second weighting, termed *streamline density with length correction* (SDL-weighted), again records streamline density but with a correction for streamline length (Hagmann et al., 2008),

$$a_{ij} = \frac{2}{g_i + g_j} \sum_{s \in S_{ij}} \frac{1}{l(s)}, \tag{2}$$

where $l(s)$ is the length of streamline s between node i and node j. The rationale for the normalisation by g (Eq. 1, 2) is to correct for between-subject variability in grey matter volume, since the number of possible entry/exit points per region is proportional to grey matter volume (Hagmann et al., 2008). The rationale for streamline length normalisation (Eq. 2) is twofold: 1) to compensate for the accumulated tractography errors which increase with streamline length; and 2) to correct for a bias in repeatedly identifying long tracts when seeding from white matter (Hagmann et al., 2008). *FA-weighted* networks were constructed from the same set of streamlines by recording the mean FA value along interconnecting streamlines. Each entry in the adjacency matrix was computed,

$$a_{ij} = \frac{1}{|S_{ij}|} \sum_{s \in S_{ij}} \frac{\sum_{v \in V_s} \mathrm{FA}(v)}{m_s}, \tag{3}$$

where V_s is the set of voxels (of size m_s) found along the streamline s between node i and node j, and FA measures the diffusion anisotropy per voxel. For each type of weighting, the result is an undirected positive-weighted graph. Self-connections were removed.

2.9. Network measures

For each $n \times n$ network, four commonly used graph-theoretic measures were computed (Rubinov and Sporns, 2010). These include two basic measures of connectivity, the *node degree*,

$$k_i = \sum_{j=1}^{n} \begin{cases} 1 & \text{if } a_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

measuring the count of links per node, and the *node strength*,

$$w_i = \sum_{j=1}^{n} a_{ij},$$

measuring the sum of edge weights per node. From each adjacency matrix, a distance matrix d was constructed recording the shortest weighted path length between any pair of nodes. A measure of integration, the average *path length* (average distance between node i and all other nodes) was computed (Watts and Strogatz, 1998),

$$l_i = \frac{1}{(n-1)} \sum_{j=1, j \neq i}^{n} d_{ij}. \tag{5}$$

Lastly, a measure of segregation, the weighted *clustering coefficient* reflecting the inter-connectivity of a node's neighbours was computed,

$$c_i = \frac{1}{k_i(k_i - 1)} \sum_{j,h=1}^{n} (a_{ij} a_{ih} a_{jh})^{1/3}. \tag{6}$$

For each of the four measures, global (nodal mean) values were also computed which were termed, *network degree*, *network strength*, *characteristic path length* and *network clustering coefficient*. Where appropriate, network properties were computed using weighted variants rather than thresholding or binarising the adjacency matrices.

2.10. Test-retest statistics

For each metric, two measures of agreement between sessions were computed, ICC (Shrout and Fleiss, 1979), and a comparison of the within- versus between-subject differences. The ICC was originally formulated for assessing multiple raters in measuring the same quantity. We computed ICC(3,1) using two-way mixed single measures using consistency of measurements between sessions, rather than absolute agreement. For a paired set of N subject-specific measures, $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_N$, the absolute *within-subject differences* were computed,

$$\delta_i^{WS} = |X_i - Y_i|. \tag{7}$$

The average *between-subject differences* (average differences of each subject against the others) were computed,

$$\delta_i^{BS} = \frac{1}{N-1} \sum_{j=1, i \neq j}^{N} |X_i - Y_j|. \tag{8}$$

Regional within- and between-subject differences were computed per node for each of the four network measures. Global within- and between-subject differences were computed from the global network properties. In each case, to test that $\delta^{WS} < \delta^{BS}$, a percentile bootstrap of the mean differences was used to compare the within- and between-subject components. The mean differences between the within- and between-subject components,

$$\delta = \frac{1}{N} \sum_{i}^{N} (\delta_i^{BS} - \delta_i^{WS}), \tag{9}$$

were computed over 5000 iterations, each time resampling with replacement from the original samples. From the distributions over these 5000 iterations, p-values were calculated as the number of times that δ was larger (or smaller) than zero. Simultaneous probability coverage and correction for multiple comparisons was obtained by adjusting the alpha level following Wilcox (2005).

3. Results

Figure 1 shows an example of segmentation, tractography and the resulting network for one subject. Visual inspection of the segmentations for each subject indicated that the FreeSurfer morphometric procedure provided plausible brain extraction, tissue segmentation and cortical labelling. Figure 2 shows the mean connectivity matrices and corresponding histograms of weights generated for the three network weightings, using the networks computed by FDT with white matter seeding for illustration. In each case the networks were produced from the same set of streamlines. Both streamline density weightings (SD and SDL) follow a similar distribution, which approximates a power law; tracking results in many low weighted connections but very few strong connections. However, due to the length correction, the SDL-weighting penalised long-range links as is evident by the down-weighting of inter-hemispheric connections (Figure 2(b)). The FA-weighting produced a markedly different distribution of edge weights, reflecting the mean diffusion anisotropy along interconnecting streamlines. Note that weights below 0.1 are absent due to the FA constraint applied in tracking.

3.1. Comparison of tractography configurations

Using FACT-GM approximately $39.7 \pm 3.0\%$ (mean $\pm$ standard deviation) of the total streamlines seeded were identified as interconnections between ROIs (using no waypoint constraint). Likewise, the interconnections identified using FACT-WM, FDT-GM and FDT-WM were $32.5 \pm 4.8\%$, $31.8 \pm 3.7\%$ and $26.2 \pm 3.7\%$ respectively. Figure 3 shows the regional network reliability measured over different combinations of tractography algorithm, seeding approach and connection weighting. In each case, the mean ICC was computed from the 84 nodes and plotted over 13 thresholds of streamline filtering by length in white matter, where zero corresponds to no length constraint. Mean ICCs of node strength were used to rate the overall nodal test-retest reliability because node strength measures the sum of weights per node. An ICC score of 1 indicates

perfect agreement between sessions, while a score of less than 0.5 may indicate poor agreement. In all cases, when comparing the same algorithm and weighting, the white matter seeding strategy outperformed the grey matter strategy. For instance, FDT-GM-SD obtained a mean ICC of 0.51, whereas FDT-WM-SD obtained 0.62. For most cases the probabilistic methods outperformed the deterministic methods, particularly for white matter seeding. However, for FA-weighting, the advantage of probabilistic over deterministic was less clear (Figure 3(c)). The best performance for both tractography algorithms was obtained when using white matter seeding and SD-weighting (mean ICC of 0.62 for FDT and 0.58 for FACT). Conversely, the poorest performance was obtained using grey matter seeding and SDL-weighting (Figure 3(b)). For many cases, the waypoint threshold offered little benefit. Only in the case of SDL-weighting (Figure 3(a)), was the mean ICC found to notably increase with waypoint length, and for this weighting the best performance was found by ensuring a white matter length >= 3.5 mm. Additionally, in all cases using GM-seeding, except FACT-GM-SDL, a waypoint threshold > 0 mm improved performance over no waypoint threshold. Overall, the best tractography configuration was found to be the FDT algorithm with white matter seeding and SD-weighting, resulting in a mean ICC of 0.62 (Figure 3(a)). This configuration was then chosen to assess network properties in the remainder of this paper.

Figure 4 shows the relationship between network properties and MCMC iterations (or number of streamlines generated) for this configuration. Figure 4(a) shows the population means of the four network measures. The mean network degree increased with the number of iterations, whereas the mean clustering coefficient and path length showed an inverse relationship with the number of iterations. The mean value of network strength was found to be variable when measured over few iterations but stabilised after approximately 40 iterations. Figure 4(b) shows the corresponding mean nodal ICCs for each network measure. The mean ICC of network strength settled to around 0.62 after approximately 10 iterations, although the other three properties showed gradual increases in reliability up to 100 iterations. However, network properties computed with 200 MCMC iterations showed no further improvement in terms of the mean ICCs (data not shown).

3.2. Network properties

Figure 5 shows the regionally computed mean values for the four network measures using FDT with white matter seeding, SD-weighting and a waypoint threshold set to 0.5mm. The absolute values of node strength, path length and clustering coefficient are not especially meaningful as they are dependent on the weighting scheme used. However, the value of one node property relative to another may offer insight into network organisation. The 95% inter-percentile range suggests that the patterns of connectivity were fairly similar across subjects. Sub-cortical structures, such as thalamus, putamen, and pallidum tend to be both highly interconnected and have strong total connection weightings. The entorhinal cortex, pars orbitalis and precentral gyrus show a high clustering coefficient, suggesting their direct nodal neighbours are also strongly inter-connected. The overall network degree was 31.5 ± 15.3, indicating that on average, each node is directly linked to nearly 32/84 of the other nodes. Overall network strength was 0.72 ± 0.7 reflecting the average sum of connection weights (normalised streamline density) per node. Network characteristic path length was 1.54 ± 0.2 reflecting the average weighted path length between all pairs of nodes. Finally, network clustering coefficient was 0.006 ± 0.003, indicating that on average a node's direct neighbours have a mean connection weighting of approximately 0.6% of the maximum weight.

3.3. Global network reliability

Table 1 shows a summary of the global test-retest analysis for FDT and white matter seeding (other tractography configurations are not reported) for each of the three network weightings. In each case, networks were constructed with the waypoint threshold set to 0.5 mm as this was closest to the best performance for all three weightings. All global network properties show good within-subject agreement with all ICCs > 0.59. In all cases, the within-subject differences (δ^{WS}) were smaller than the between-subject differences (δ^{BS}). For SD-weighting, ICCs were between 0.62 and 0.76, within-subject differences 3.2 - 11.9%, between-subject differences 5.8 - 20% and the corresponding p-values between 0.018 and 0.11. For SDL-weighting, ICCs were between 0.59 and 0.76, within-subject differences 3.1 - 8.9%, between-subject differences 5.5 - 16.2% and the corresponding p-values between 0.006 and 0.11. For FA-weighting, ICCs were between 0.64 and

0.69, within-subject differences 4.5 - 9.8%, between-subject differences 8.3 - 17.5% and the corresponding p-values between 0.06 and 0.11. Overall, the global test-retest properties were broadly similar over the three weightings, although SDL-weighting obtained marginally better performance than the other weightings. For network strength and clustering coefficient the SDL-weighting produced the most significant differences, whereas for characteristic path length, the FA-weighting came closest to significance. In each case the results for network degree were identical because the same set of streamlines was used.

3.4. Regional network reliability

Table 2 shows a summary of the regional test-retest analysis. For SD-weighting, mean ICCs were between 0.46 and 0.62, mean within-subject differences 5.2 - 24.2%, mean between-subject differences 8.1 - 35.7%. For SDL-weighting, mean ICCs were between 0.45 and 0.58, mean within-subject differences 4.9 - 22.9% and mean between-subject differences 7.6 - 31.4%. For FA-weighting, mean ICCs were between 0.50 and 0.56, mean within-subject differences 6.4 - 19.0% and mean between-subject differences 10.0 - 27.3%. For all four measures, SD-weighting obtained marginally better regional test-retest performance than SDL-weighting in terms of mean ICCs. However, in the case of path length and clustering coefficient the FA-weighting obtained better performance than SD-weighting. Again, the results for node degree were identical across weightings as the same set of streamlines was used.

Figure 6 shows the regional ICCs for each of the 84 nodes, again using FDT with white matter seeding, SD-weighting and a waypoint threshold set to 0.5mm. Only 22.6% (19/24) of nodes obtained ICCs ≥ 0.5 across all four measures. 77.4% (65/84) of nodes showed poor within-subject agreement with an ICC < 0.5 across one or more measures. Furthermore, 10.7% (9/84) of nodes showed a negative ICC across one or more measures. The theoretical value of ICC is non-negative, but estimates can be negative. These nine nodes were left/right pallidum, left/right amygdala, left caudal middle frontal, left pars orbitalis, right inferior temporal, right pars opercularis and right superior parietal. Figure 7 shows the corresponding differences between the within- and between-subject components, $\delta^{BS} - \delta^{WS}$. Although, for 83.3% (70/84) of nodes the within-subject differences were smaller than the between-subject differences across all four measures, on average less than 1/8 nodes show that the differences were significant ($p < 0.05$) for our sample. Negative differences occurred for 15.5% (13/84) of nodes in at least one measure, indicating that the within-subject difference outweighed the estimated between-subject difference. These 13 nodes were right pallidum, left hippocampus, left amygdala, left caudal anterior cingulate, left caudal middle frontal, right fusiform, left lateral occipital, right lateral orbitofrontal, left parahippocampal, left/right pars orbitalis, right pericalcarine and right transverse temporal.

3.5. Variation between cortical and sub-cortical nodes

Our analysis revealed a variation in regional test-retest scores between cortical and sub-cortical nodes. For cortical nodes the mean within-subject difference (of node strength) was 24.96%, the mean between-subject difference was 37.16% and the mean ICC was 0.65. Whereas, for sub-cortical nodes the mean within-subject difference was 20.82%, the mean between-subject was 29.45% and the mean ICC was 0.50. Subsequently, we compared the 84×84 whole-brain networks with 68×68 networks constructed from cortico-cortical connections only. A percentile bootstrap contrast, which compared the mean within- versus between-subject differences (δ) between the whole-brain and cortico-cortical networks, suggested that there were no significant differences for any of the four node measures. This indicates that for our sample, the inclusion of sub-cortical connections does not reduce test-retest reliability.

4. Discussion

Preceding our work, several studies have assessed aspects of network reliability using repeat scans of healthy human volunteers. Hagmann et al. (2008) assessed structural networks obtained from diffusion spectrum imaging (DSI), while Vaessen et al. (2010) assessed reproducibility over different sets of diffusion gradient directions using diffusion tensor imaging (DTI). Bassett et al. (2010) compared reliability in both DTI and DSI, and Cammoun et al. (2011) investigated the effect of network resolution using DSI. Finally,

Cheng et al. (2012) assessed test-retest reliability using DTI with focus on the differences between binary and weighted networks. These studies have addressed network reliability in different ways. Hagmann et al. found a within-subject network correlation of 0.78 (N=1) and a mean between-subject correlation of 0.65 (N=5). Similarly, Cammoun et al. compared matrices directly by Pearson's correlation coefficient and found within-subject correlations ranging from 0.874 to 0.976 (N=5) and between-subject correlations of between 0.724 to 0.958 (N=20) across five different network resolutions. Vaessen et al. found within-subject coefficients of variation (CV) $< 3.8\%$ and ICCs between 0 and 0.94 for node degree, path length and clustering coefficient (N=6). Bassett et al. found within-subject CVs $< 5\%$ and ICCs > 0.72 (N=7). Cheng et al. found good test-retest agreement in both global and regional network properties (N=44). Though it is not straightforward to compare between assays, these studies indicate that, globally, same-subject networks differ between scanning sessions yet the between-subject variation is typically greater than the within-subject variation.

We obtained similar results in assessing global network properties. Using probabilistic tractography and white matter seeding, the global within-subject differences ($< 11.9\%$) were smaller than the global between-subject differences with p-values at either $p < 0.05$ or trend level and ICCs > 0.59 (Table 1). These findings indicate that global network properties can be estimated reliably from session to session for all three types of network weighting tested. Overall, the network weighting recording streamline density with streamline length correction (Hagmann et al., 2008) showed marginally better global test-retest performance that the other two weightings.

The comparison of tractography algorithms, seeding approaches and network weightings (Figure 3) offers some insights into network reliability. Firstly, in all cases white matter seeding produced networks with better test-retest reliability than grey matter seeding, in terms of the mean ICC measuring node strength. We note that these results may be partly biased due to the larger sample obtained with white matter seeding, as our setup involves seeding from all voxels in either grey or white matter, and there are more seed points in white matter than grey. Nevertheless, we believe some grey matter seeding error may arise because all grey matter voxels were included even those unlikely to be involved in interfacing with white matter, potentially resulting in a greater proportion of spurious tracts. Investigation of other seeding approaches merits further investigation, notably, approaches which seed from both grey and white matter (whole-brain) and approaches which seed only from the interface of the grey-white matter (see Robinson et al. 2010; Vaessen et al. 2010), which may exclude a proportion of inappropriate seed points. Secondly, overall, the probabilistic algorithm produced better test-retest performance than the deterministic method. This is perhaps due to the limited sample of possible streamlines produced by the deterministic method. The deterministic approach estimates a best fit of the diffusion tensor model at each voxel, whereas the probabilistic approach estimates a distribution. Concerning probabilistic tractography, the fibre model and the number of iterations are important considerations. The general recommendation is to use 1000 or more iterations per seed point. In this study, we applied 100 streamlines per seed voxel, equating to approximately 6 million streamlines per subject. However, the results obtained when doubling the number of iterations to 200 showed no further increase in agreement between sessions (data not shown). Note that these conclusions were drawn from the mean ICC scores (Figure 4(b)) and further analysis may be required to test this point. Unfortunately, networks constructed with thousands of MCMC iterations are often impractical due to the computational cost involved.

For white matter seeding, the regional findings (Figure 3) showed that the white matter waypoint constraints were largely ineffective in reducing false connections, except in the case of SDL-weighting. However, for grey matter seeding the waypoint constraints improved the regional network reliability for all three types of weighting in comparison to unrestricted connectivity mapping. This is because grey matter seeding without any streamline filtering can produce streamlines which connect directly to a neighbouring grey matter ROI without apparently passing through any white matter. Due to partial volume effects some of these connections may be genuine, particularly at the grey-white matter interface, but this is challenging to validate at dMRI resolution and such connections may often be spurious. However, we believe a constraint on white matter length is a more valid method of streamline filtering than thresholding purely by length. Waypoint constraints use prior knowledge of the white matter location to filter streamlines, whereas arbitrary length thresholds may remove genuine streamlines regardless of whether they pass through white matter. Our

results indicate that constraining streamlines to pass through at least one white matter voxel was enough to improve network reliability.

An important consideration when quantifying interconnections is the normalisation applied to correct streamline counts for effects due to differences in grey and/or white matter volumes. However, some researchers have suggested that volume correction (e.g. SD and SDL-weightings) may overcompensate for volume-driven effects on streamline counts (Van Den Heuvel and Sporns, 2011). It is not clear how such effects may be counteracted to allow representative comparison of connectivity between individuals. Alternatively, instead of streamline density it is possible to record some measure of tissue microstructure, such as diffusion anisotropy, averaged along the length of each tract (Iturria-Medina et al., 2007; Robinson et al., 2010). Such weightings somewhat circumvent the need for correction, as the weights reflect the underlying diffusion properties and are thus less affected by differences in tissue volume. Although, we found that FA-weighting provided poorer test-retest performance than the SD-weighting, it is possible that network weights based on properties of tissue microstructure may be more appropriate for group-wise analysis.

Some researchers have pointed out a bias arising from the tracking procedure, meaning that the number of fibres identified between a pair of regions decreases as a function of the distance (Zalesky and Fornito, 2009). This is due to the greater number of propagation steps and hence accumulated error associated with longer streamlines. An attempt to correct such biases may be made by a normalisation on streamline length. We compared networks with and without length correction (SDL and SD-weighting) finding that in almost all cases, whether seeding from grey or white matter, that the uncorrected weighting (SD) produced networks with better regional reliability (Figure 3), although SDL obtained marginally superior global reliability (Table 1). This suggests that for our data, the length correction overcompensates for accumulated errors in longer streamlines. Cheng et al. (2012) made similar findings, suggesting that the length correction may not be necessary because the success rate for inter-region tracks is lower with longer fibres.

In this study, we have chosen to avoid thresholding weights from each connectivity matrix due to concerns that such thresholds may bias the results of a group-wise analysis. It is difficult to ensure that an arbitrary threshold removes spurious connections while retaining genuine patterns of connectivity. Additionally, applying the same threshold to multiple connectivity matrices is likely to results in different levels of sparsity, whereas matching sparsity across subjects is also problematic (Zalesky et al., 2010a). Nevertheless, trials comparing a range of suprathresholds on the connection weights indicated that some thresholds offered improvement in the overall test-retest reliability of connection (edge) statistics but marginal improvements in nodal and global properties (data not shown). However, such thresholds should be used with caution. In this work, we have pursued anatomical constraints in order to reduce spurious connections, such as the FA, curvature and waypoint thresholds used in tractography.

Overall, our findings highlight some concerns about the regional reliability of dMRI networks, suggesting that the connections to some nodes are computed unreliably from session to session (Table 2, Figure 6, 7). On average, only one in eight nodes show a significant difference in between- and within-subject variation for our sample. In addition, only 22.6% (19/84) nodes have ICCs ≥ 0.5 across all four measures.

4.1. Sources of test-retest variation

Some test-retest variation may be due to scanner noise and inhomogeneities between sessions and some may be due to systematic variation in processing (Hagmann et al., 2010; Van Essen and Ugurbil, 2012). Evidently, many intermediate steps are involved in generating structural networks from dMRI data. The variability at each step contributes to the variability in the following stage and in the resulting measures of connectivity. In particular, following our approach, network properties are dependent on reliable registration, node segmentation, diffusion processing and tractography.

In our case, the regional network properties are reliant on the FreeSurfer morphometric procedure. These methods have been demonstrated to show good test-retest reliability across scanner manufacturers and across field strengths (Han et al., 2006; Wonderlick et al., 2009; Reuter et al., 2012), although other studies have shown discrepancies in test-retest reliability (Morey et al., 2010; Gronenschild et al., 2012). It should be noted that the task of automated cortical labelling is extremely challenging.

An appropriate definition of network nodes is essential but far from trivial (Hagmann et al., 2010; Zalesky et al., 2010b). A network formed from too few large nodes may fail to capture the true connectivity between

regions. However, a network formed from too many small nodes may be influenced by systematic errors and noise, leading to spurious findings. Ideally, we wish to obtain a network resolution where the measurement of genuine structural differences is larger than the noise measurements. We chose the Desikan-Killiany atlas for our analysis as using larger and fewer nodes may minimise the effects of image noise and systematic error. We also assessed a 164 node parcellation (Destrieux et al., 2010) but test-retest results were poorer than with the 84 node configuration and therefore are not reported. Given the resolution used in this study (84 nodes), our findings suggest that the measurement noise may degrade some of the genuine patterns of connectivity for several nodes. However, networks of 50-100 nodes are typical for current connectome studies. Additionally, many studies have identified cortico-cortical connections only, even though, sub-cortical structures have an essential role in brain wiring. For instance, the thalamus is highly connected to many cortical regions (Behrens et al., 2003a). Therefore, we found it important to include sub-cortical structures in our analysis. Although we found a variation in the nodal properties between cortical and sub-cortical regions, the differences between whole-brain and cortico-cortical networks were not significant (Section 3.5).

Additionally, some error may reflect tractography issues in estimating the underlying axonal tracts. This may due to both ROI segmentation errors affecting seeding and methodological issues in streamline construction. Validation of tractography has been performed with good agreement with the underlying neuronal connections in the porcine and macaque brain (Parker et al., 2002; Dyrby et al., 2007). However, tractography is know to be strongly affected by measurement noise resulting in both false positive and false negative connections (Jbabdi and Johansen-Berg, 2011; Zalesky and Fornito, 2009). Additionally, one cannot expect to map fibre bundles smaller than the dMRI resolution and there are issues accounting for crossing, branching and kissing fibres. See Yo et al. (2009) for a comparison of various tractography algorithms in a structural connectome context. The uncertainty in fibre directions for noisy measurements may also be a factor in the poor test-retest results. Fillard et al. (2011) suggested that for medium or low signal-to-noise datasets, an appropriate prior on the spatial smoothness of either the diffusion model or the fibres is recommended for correct modelling.

4.2. Limitations of study

Analysis of reliability is challenging because results depend on both the variables analysed and the metric used. Here we used four dependant variables: node degree, node strength, path length and clustering coefficient. There are of course many other ways to characterise networks but these measures are the basis of graph-theoretic analyses. It is worth noting that in this study agreement was based on either global or nodal measures, which are themselves a function of elements of the adjacency matrix. It is also possible to directly assess every (non-zero) element of the adjacency matrix (see Zalesky et al. 2010a), which may offer further insight into network reliability. In terms of metrics, we relied on both the within- versus between-subject differences and the ICC. One issue with the ICC is that large samples may be required to estimate scores to acceptable precision. Shoukri et al. (2004) determined that for 2 repeated measures in order to estimate a minimum acceptable ICC score of 0.8 with 95% confidence intervals of width 0.2, then 52 subjects are required (see Table 3 Shoukri et al. 2004). Similarly, to estimate a minimum acceptable ICC score of 0.6 with 95% confidence intervals of width 0.2, then 158 subjects are required. Clearly, such samples may be unrealistic in the case of MRI, and to date, no test-retest study has assessed a sample of this size, although Cheng et al. 2012 used 44 subjects. Therefore, the studies relying on ICC (and also Pearson's correlation coefficient, which itself is a biased version of ICC because it assumes independent variances) have biased estimates. We can thus recommend for future dMRI connectome studies to use a sample of more than 50 subjects. Furthermore, as well as ICCs we have also quantified test-retest performance by comparing the within-subject to the mean between-subject differences. For a small sample the measurement error may be quite high, but bootstrap estimates are less affected by sample size. For most nodes the within-subject differences were smaller than between-subject differences yet most differences were not found to be significant, which can also reflect a lack of power. Nevertheless, our analysis provides insight into whether the genuine patterns of connectivity can be identified despite noisy measurements.

5. Conclusion

We constructed networks from structural MRI and dMRI data obtained from ten healthy volunteers scanned on two separate occasions. The subjects had a narrow age range (50-58 years) to minimise the possible confound of increasing age on connectivity and diffusion anisotropy values. Network reliability was assessed by varying a number of factors affecting network construction. Our findings suggest that test-retest performance was improved when: 1) seeding from white matter, rather than grey; and 2) using probabilistic tractography with a two-fibre model and sufficient streamlines, rather than deterministic tensor tractography. However, a potential strategy to reduce false connections based on streamline length was largely ineffective in improving reliability. In terms of network weighting, a measure of streamline density produced better test-retest performance than tract-averaged FA, although it remains unclear which is a more accurate representation of the underlying connectivity. Our findings suggest that current connectome mapping techniques (at 1.5 T) are adequate for reliably measuring global network measures. However, regional network measures may not be as reliable, leading to concerns about the validity of studies based on such measures, particularly with small sample sizes. Future work should be undertaken to address these concerns.

Conflict of Interest Statement

All authors declare no conflicts of interest.

Acknowledgements

All imaging was performed in the Brain Research Imaging Centre, University of Edinburgh (`http://www.bric.ed.ac.uk`), and was funded by the Edinburgh Experimental Cancer Medicine Centre. CB was funded by the UK Engineering and Physical Sciences Research Council and The Medical Research Council through the Doctoral Training Centre in Neuroinformatics and Computational Neuroscience, University of Edinburgh. CP is partly funded by SINAPSE (Scottish Imaging Network: A Platform for Scientific Excellence; `http://www.sinapse.ac.uk`).

References

Andersson, J.L.R., Jenkinson, M., Smith, S., 2007. Non-linear registration aka Spatial normalisation. Technical Report TR07JA2. Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, University of Oxford.

Basser, P.J., Pajevic, S., Pierpaoli, C., Duda, J., Aldroubi, A., 2000. In vivo fiber tractography using DT-MRI data. Magnetic Resonance in Medicine 44, 625–32.

Basser, P.J., Pierpaoli, C., 1996. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. Journal of magnetic resonance Series B 111, 209–219.

Bassett, D.S., Brown, J.A., Deshpande, V., Carlson, J.M., Grafton, S.T., 2010. Conserved and variable architecture of human white matter connectivity. NeuroImage 54, 1262–1279.

Behrens, T.E.J., Berg, H.J., Jbabdi, S., Rushworth, M.F.S., Woolrich, M.W., 2007. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? NeuroImage 34, 144–155.

Behrens, T.E.J., Johansen-Berg, H., Woolrich, M.W., Smith, S.M., Wheeler-Kingshott, C.A.M., Boulby, P.A., Barker, G.J., Sillery, E.L., Sheehan, K., Ciccarelli, O., Thompson, A.J., Brady, J.M., Matthews, P.M., 2003a. Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. Nature Neuroscience 6, 750–757.

Behrens, T.E.J., Woolrich, M.W., Jenkinson, M., Johansen-Berg, H., Nunes, R.G., Clare, S., Matthews, P.M., Brady, J.M., Smith, S.M., 2003b. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. Magnetic Resonance in Medicine 50, 1077–1088.

Cammoun, L., Gigandet, X., Meskaldji, D., Thiran, J.P., Sporns, O., Do, K.Q., Maeder, P., Meuli, R., Hagmann, P., 2011. Mapping the human connectome at multiple scales with diffusion spectrum MRI. Journal of Neuroscience Methods 6, 1–12.

Cheng, H., Wang, Y., Sheng, J., Kronenberger, W.G., Mathews, V.P., Hummer, T.A., Saykin, A.J., 2012. Characteristics and variability of structural networks derived from diffusion tensor imaging. NeuroImage 61, 1153–64.

Cook, P.A., Alexander, D.C., 2006. Modelling uncertainty in two fibre-orientation estimates within a voxel. Methods 48, 2006–2006.

Crofts, J.J., Higham, D.J., Bosnell, R., Jbabdi, S., Matthews, P.M., Behrens, T.E.J., Johansen-Berg, H., 2010. Network analysis detects changes in the contralesional hemisphere following stroke. NeuroImage 54, 161–169.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980.

Destrieux, C., Fischl, B., Dale, A., Halgren, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53, 1–15.

Dyrby, T.B., Sø gaard, L.V., Parker, G.J., Alexander, D.C., Lind, N.M., Baaré, W.F.C., Hay-Schmidt, A., Eriksen, N., Pakkenberg, B., Paulson, O.B., Jelsing, J., 2007. Validation of in vitro probabilistic tractography. NeuroImage 37, 1267–1277.

Fillard, P., Descoteaux, M., Goh, A., Gouttard, S., Jeurissen, B., Malcolm, J., Ramirez-Manzanares, A., Reisert, M., Sakaie, K., Tensaouti, F., Yo, T., Mangin, J.F., Poupon, C., 2011. Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. NeuroImage 56, 220–234.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Kouwe, A.V.D., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole Brain Segmentation: Neurotechnique Automated Labeling of Neuroanatomical Structures in the Human Brain. Neuron 33, 341–355.

Fischl, B., Salat, D.H., Van Der Kouwe, A.J.W., Makris, N., Ségonne, F., Quinn, B.T., Dale, A.M., 2004a. Sequence-independent segmentation of magnetic resonance images. NeuroImage 23 Suppl 1, S69–84.

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, A.M., 2004b. Automatically parcellating the human cerebral cortex. Cerebral Cortex 14, 11–22.

Gong, G., Rosa-Neto, P., Carbonell, F., Chen, Z.J., He, Y., Evans, A.C., 2009. Age- and gender-related differences in the cortical anatomical network. Journal of Neuroscience 29, 15684–15693.

Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., Ghosh, S.S., 2011. Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. Frontiers in neuroinformatics 5, 15.

Gorgolewski, K.J., Storkey, A., Bastin, M.E., Whittle, I.R., Wardlaw, J.M., Pernet, C.R., 2013. A test-retest fMRI dataset for motor, language and spatial attention functions. GigaScience 2, 6.

Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. NeuroImage 48, 63–72.

Gronenschild, E.H.B.M., Habets, P., Jacobs, H.I.L., Mengelers, R., Rozendaal, N., Van Os, J., Marcelis, M., 2012. The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. PLoS ONE 7, e38234.

Hagmann, P., 2005. From diffusion MRI to brain connectomics. Ph.D. thesis. ITS Institut de traitement des signaux.

Hagmann, P., Cammoun, L., Gigandet, X., Gerhard, S., Ellen Grant, P., Wedeen, V., Meuli, R., Thiran, J.P., Honey, C.J., Sporns, O., 2010. MR connectomics: Principles and challenges. J Neurosci Methods 194, 34–45.

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C.J., Wedeen, V.J., Sporns, O., 2008. Mapping the Structural Core of Human Cerebral Cortex. PLoS Biology 6, 15.

Hagmann, P., Kurant, M., Gigandet, X., Thiran, P., Wedeen, V.J., Meuli, R., Thiran, J.P., 2007. Mapping Human Whole-Brain Structural Networks with Diffusion MRI. PLoS ONE 2, 9.

Han, X., Jovicich, J., Salat, D., Van Der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R.,

Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., Fischl, B., 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. NeuroImage 32, 180–194.
Honey, C., Sporns, O., Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., 2008. The Structural Core of Human Cerebral Cortex and its relation to the brains default network, in: Proceedings 16th Scientific Meeting International Society for Magnetic Resonance in Medicine, p. 839.
Iturria-Medina, Y., Canales-Rodríguez, E.J., Melie-García, L., Valdés-Hernández, P.A., Martínez-Montes, E., Alemán-Gómez, Y., Sánchez-Bornot, J.M., 2007. Characterizing brain anatomical connections using diffusion weighted MRI and graph theory. NeuroImage 36, 645–660.
Jbabdi, S., Johansen-Berg, H., 2011. Tractography: where do we go from here? Brain Connectivity 1, 169–183.
Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.
Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Medical Image Analysis 5, 143–156.
Jones, D.K., Williams, S.C.R., Gasston, D., Horsfield, M.A., Simmons, A., Howard, R., 2002. Isotropic resolution diffusion tensor imaging with whole brain acquisition in a clinically acceptable time. Human brain mapping 15, 216–230.
Lazar, M., Weinstein, D.M., Tsuruda, J.S., Hasan, K.M., Arfanakis, K., Meyerand, M.E., Badie, B., Rowley, H.A., Haughton, V., Field, A., Alexander, A.L., 2003. White matter tractography using diffusion tensor deflection. Human Brain Mapping 18, 306–321.
Lo, C.Y., Wang, P.N., Chou, K.H., Wang, J., He, Y., Lin, C.P., 2010. Diffusion tensor tractography reveals abnormal topological organization in structural cortical networks in Alzheimer's disease. Journal of Neuroscience 30, 16876–16885.
Maldjian, J.A., Laurienti, P.J., Kraft, R.A., Burdette, J.H., 2003. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. NeuroImage 19, 1233–1239.
Morey, R.A., Selgrade, E.S., Wagner, H.R., Huettel, S.A., Wang, L., McCarthy, G., 2010. Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. Human brain mapping 31, 1751–1762.
Mori, S., Crain, B.J., Chacko, V.P., Van Zijl, P.C., 1999. Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. Annals of Neurology 45, 265–269.
Parker, G.J.M., Haroon, H.A., Wheeler-Kingshott, C.A.M., 2003. A framework for a streamline-based probabilistic index of connectivity (PICo) using a structural interpretation of MRI diffusion measurements. Journal of Magnetic Resonance Imaging 18, 242–254.
Parker, G.J.M., Stephan, K.E., Barker, G.J., Rowe, J.B., MacManus, D.G., Wheeler-Kingshott, C.A.M., Ciccarelli, O., Passingham, R.E., Spinks, R.L., Lemon, R.N., Turner, R., 2002. Initial demonstration of in vivo tracing of axonal projections in the macaque brain and comparison with the human brain using diffusion tensor Imaging and fast marching tractography. NeuroImage 15, 797–809.
Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. NeuroImage 61, 1402–1418.
Robinson, E.C., Hammers, A., Ericsson, A., Edwards, A.D., Rueckert, D., 2010. Identifying population differences in whole-brain structural networks: a machine learning approach. NeuroImage 50, 910–919.
Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: uses and interpretations. NeuroImage 52, 1059–1069.
Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W., 2008. Construction of a 3D probabilistic atlas of human cortical structures. NeuroImage 39, 1064–1080.
Shoukri, M.M., Asyali, M.H., Donner, A., 2004. Sample size requirements for the design of reliability study: review and new results. Statistical Methods in Medical Research 13, 251–271.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86, 420–428.
Shu, N., Liu, Y., Li, K., Duan, Y., Wang, J., Yu, C., Dong, H., Ye, J., He, Y., 2011. Diffusion Tensor Tractography Reveals Disrupted Topological Efficiency in White Matter Structural Networks in Multiple Sclerosis. Cerebral Cortex 21, 2565–2577.
Skudlarski, P., Jagannathan, K., Anderson, K., Stevens, M.C., Calhoun, V.D., Skudlarska, B.A., Pearlson, G., 2010. Brain connectivity is not only lower but different in schizophrenia: a combined anatomical and functional approach. Biological Psychiatry 68, 61–69.
Smith, S.M., 2002. Fast robust automated brain extraction. Human Brain Mapping 17, 143–155.
Sporns, O., 2011. The human connectome: a complex network. Annals Of The New York Academy Of Sciences 1224, 109–125.
Sporns, O., Tononi, G., Kötter, R., 2005. The Human Connectome: A Structural Description of the Human Brain. PLoS Computational Biology 1, e42.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15, 273–289.
Vaessen, M.J., Hofman, P.A.M., Tijssen, H.N., Aldenkamp, A.P., Jansen, J.F.A., Backes, W.H., 2010. The effect and reproducibility of different clinical DTI gradient sets on small world brain connectivity measures. NeuroImage 51, 1106–1116.
Van Den Heuvel, M.P., Sporns, O., 2011. Rich-Club Organization of the Human Connectome. Journal of Neuroscience 31, 15775–15786.
Van Essen, D.C., Ugurbil, K., 2012. The future of the human connectome. NeuroImage 62, 1–12.
Verstraete, E., Veldink, J.H., Mandl, R.C.W., Van Den Berg, L.H., Van Den Heuvel, M.P., 2011. Impaired Structural Motor Connectome in Amyotrophic Lateral Sclerosis. PLoS ONE 6, 10.
Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393, 440–442.
Wee, C.Y., Yap, P.T., Li, W., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D., 2011.

Enriched white matter connectivity networks for accurate identification of MCI patients. NeuroImage 54, 1812–1822.
Wen, W., Zhu, W., He, Y., Kochan, N.A., Reppermund, S., Slavin, M.J., Brodaty, H., Crawford, J., Xia, A., Sachdev, P., 2011. Discrete neuroanatomical networks are associated with specific cognitive abilities in old age. Journal of Neuroscience 31, 1204–1212.
Wilcox, R.R., 2005. Introduction to robust estimation and hypothesis testing. Second ed., Academic Press.
Wonderlick, J.S., Ziegler, D.A., Hosseini-Varnamkhasti, P., Locascio, J.J., Bakkour, A., Van Der Kouwe, A., Triantafyllou, C., Corkin, S., Dickerson, B.C., 2009. Reliability of MRI-derived cortical and subcortical morphometric measures: effects of pulse sequence, voxel geometry, and parallel imaging. NeuroImage 44, 1324–1333.
Yan, C., Gong, G., Wang, J., Wang, D., Liu, D., Zhu, C., Chen, Z.J., Evans, A., Zang, Y., He, Y., 2011. Sex- and brain size-related small-world structural cortical networks in young adults: a DTI tractography study. Cerebral Cortex 21, 449–458.
Yo, T.S., Anwander, A., Descoteaux, M., Fillard, P., Poupon, C., Knösche, T.R., 2009. Quantifying brain connectivity: a comparative tractography study. Medical Image Computing and Computer-Assisted Intervention 12, 886–893.
Ystad, M., Hodneland, E., Adolfsdottir, S., Haász, J., Lundervold, A.J., Eichele, T., Lundervold, A., 2011. Cortico-striatal connectivity and cognition in normal aging: a combined DTI and resting state fMRI study. NeuroImage 55, 24–31.
Zalesky, A., Fornito, A., 2009. A DTI-derived measure of cortico-cortical connectivity. IEEE Transactions on Medical Imaging 28, 1023–1036.
Zalesky, A., Fornito, A., Bullmore, E.T., 2010a. Network-based statistic: identifying differences in brain networks. NeuroImage 53, 1197–1207.
Zalesky, A., Fornito, A., Harding, I.H., Cocchi, L., Yücel, M., Pantelis, C., Bullmore, E.T., 2010b. Whole-brain anatomical networks: does the choice of nodes matter? NeuroImage 50, 970–983.
Zalesky, A., Fornito, A., Seal, M.L., Cocchi, L., Westin, C.F., Bullmore, E.T., Egan, G.F., Pantelis, C., 2011. Disrupted axonal fiber connectivity in schizophrenia. Biological Psychiatry 69, 80–89.

Figures

Figure 1: (A) FreeSurfer cortical parcellation visualised on pial surface (56 year old male); (B) Interconnecting streamlines constructed by probabilistic tractography, filtered by curvature and length for visualisation, where colour indicates the x, y, z (red, green, blue) direction of each streamline segment; (C) Graph representation of the resulting structural network, where node size indicates node strength and edge width indicates connection weight.

Figure 2: Top row: 84×84 mean connectivity matrices of inter-region connections averaged across all subjects (N=20) for the three network weightings and generated from the same set of streamlines: (A) streamline density; (B) streamline density with streamline length correction, (C) tract-averaged FA. Note, that both SD and SDL-weighted matrices were scaled between zero and one and (A) and (B) are log scaled. In each case, the two large rectangular patterns on the diagonal correspond to the left and right hemispheres and the node ordering is the same as Figures 5-7. Bottom row: the corresponding histograms of the connection weights for each of the three network weightings.

Figure 3: Mean ICC values measuring test-retest reliability of node strength over thirteen thresholds of streamline filtering by white matter waypoint length for each tractography algorithm, seeding type and network weighting. 95% confidence intervals of the mean were estimated by resampling with replacement 5000 times from the 84 nodes and recomputing the mean and taking the 2.5 and 97.5 percentiles of this distribution.

Figure 4: The relationship between network properties and number of MCMC iterations using FDT with white matter seeding, streamline density weighting and waypoint threshold set to 0.5 mm: (A) The population means of the four network measures; (B) The corresponding mean regional ICCs for each network measure.

Figure 5: Nodal values computed from population (N=20) means, from top-to-bottom: node degree, node strength, path length and clustering coefficient. Error bars show the 95% inter-percentile range.

Figure 6: Nodal ICC values with 95% confidence intervals (N=10) from top-to-bottom: node degree, node strength, path length and clustering coefficient. Negative correlations are not shown.

Figure 7: Mean nodal differences ($\delta^{BS} - \delta^{WS}$) with 95% confidence intervals of the mean estimated by bootstrap resampling (N=10), from top-to-bottom: node degree, node strength, path length and clustering coefficient. Negative differences are not shown.

FDT, WM-seeding, SD-weighted

	ICC	mean δ^{WS} %		mean δ^{BS} %		p-value
network degree	0.66	8.22	(8.3)	14.10	(11.0)	0.110
network strength	0.75	11.93	(9.9)	19.95	(14.4)	0.066
characteristic path length	0.62	3.22	(3.6)	5.83	(4.8)	0.099
network clustering coefficient	0.76	6.37	(6.2)	12.56	(9.3)	**0.018**

FDT, WM-seeding, SDL-weighted

	ICC	mean δ^{WS} %		mean δ^{BS} %		p-value
network degree	0.66	8.22	(8.3)	14.10	(11.0)	0.110
network strength	0.76	8.86	(6.1)	15.09	(10.5)	**0.025**
characteristic path length	0.59	3.06	(3.3)	5.51	(4.5)	0.097
network clustering coefficient	0.71	8.35	(7.2)	16.18	(9.6)	**0.006**

FDT, WM-seeding, FA-weighted

	ICC	mean δ^{WS} %		mean δ^{BS} %		p-value
network degree	0.66	8.22	(8.3)	14.10	(11.0)	0.110
network strength	0.67	9.82	(10.5)	17.54	(13.9)	0.070
characteristic path length	0.64	4.46	(4.8)	8.25	(7.1)	0.060
network clustering coefficient	0.69	4.77	(5.3)	8.87	(8.5)	0.073

Table 1: Summary of the global test-retest analysis for the three types of network weighting using FDT and white matter seeding, showing: ICC, mean within-subject differences (δ^{WS}), mean between-subject differences (δ^{BS}). The differences are expressed as a percentage, bracketing indicates the standard deviation and emboldening indicates significance ($p < 0.05$).

FDT, WM-seeding, SD-weighted

	mean ICC		mean δ^{WS} %		mean δ^{BS} %	
node degree	0.50	(0.3)	16.35	(7.0)	23.25	(8.1)
node strength	0.62	(0.2)	24.18	(8.3)	35.69	(11.0)
path length	0.53	(0.2)	5.22	(1.5)	8.08	(2.0)
clustering coefficient	0.46	(0.3)	21.82	(7.4)	28.77	(8.8)

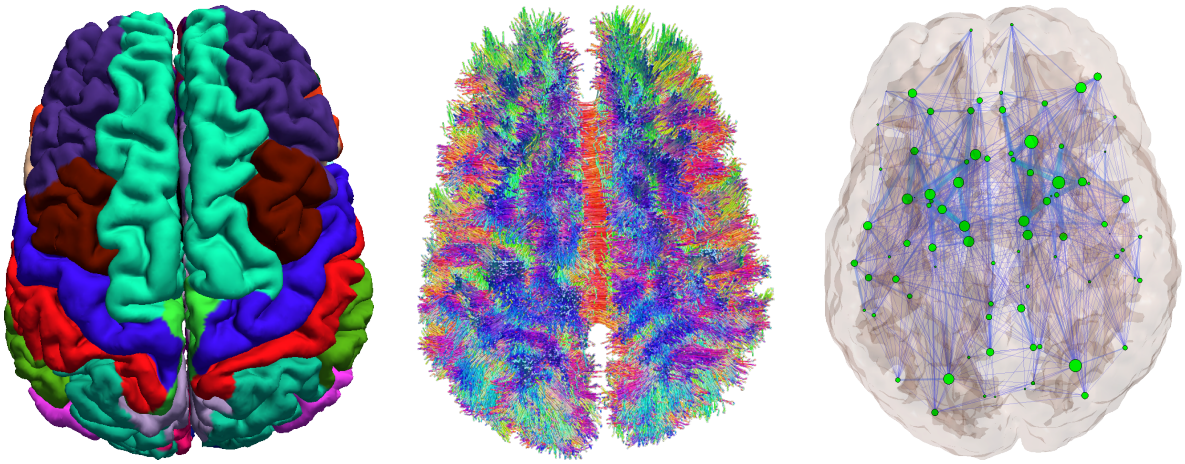
FDT, WM-seeding, SDL-weighted

	mean ICC		mean δ^{WS} %		mean δ^{BS} %	
node degree	0.50	(0.3)	16.35	(7.0)	23.25	(8.1)
node strength	0.58	(0.2)	21.85	(7.7)	31.11	(10.5)
path length	0.51	(0.3)	4.91	(1.5)	7.57	(2.0)
clustering coefficient	0.45	(0.3)	22.92	(8.8)	31.36	(9.8)

FDT, WM-seeding, FA-weighted

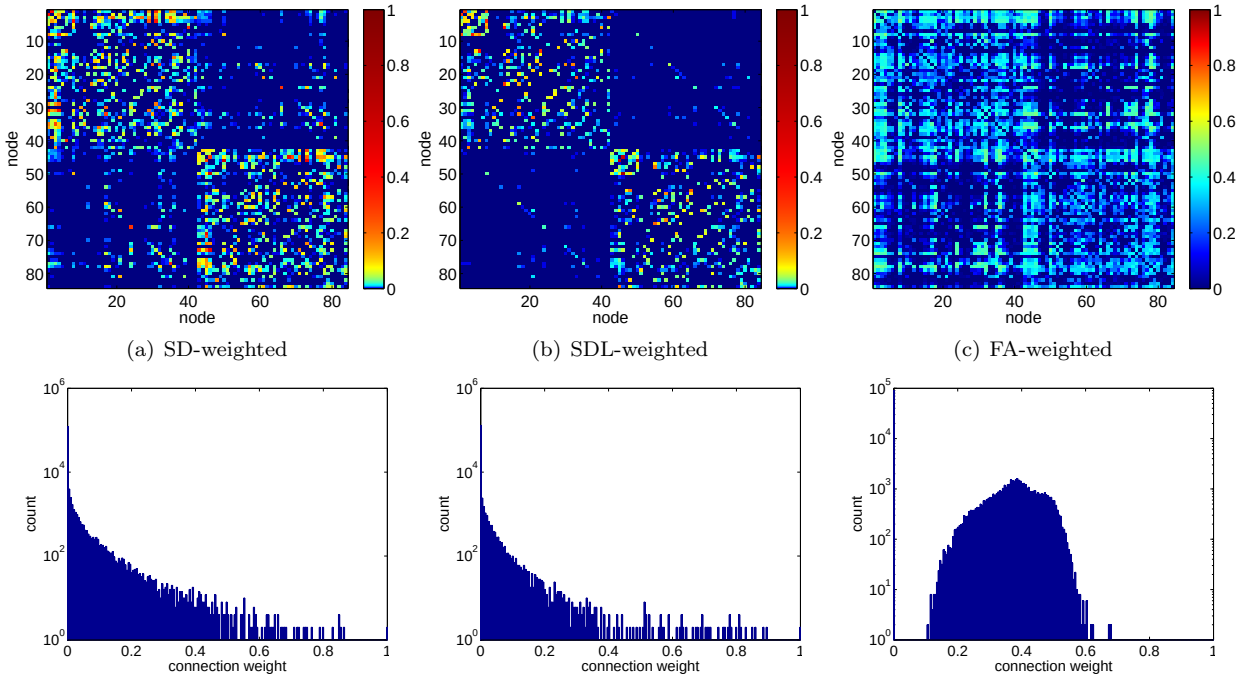
	mean ICC		mean δ^{WS} %		mean δ^{BS} %	
node degree	0.50	(0.3)	16.35	(7.0)	23.25	(8.1)
node strength	0.52	(0.3)	19.02	(8.0)	27.28	(9.2)
path length	0.56	(0.2)	6.35	(1.9)	10.02	(2.4)
clustering coefficient	0.51	(0.2)	8.21	(2.2)	12.38	(2.8)

Table 2: Summary of the regional (nodal) test-retest analysis for all 84 nodes for the three types of network weighting using FDT and white matter seeding, showing: mean ICC, mean within-subject differences (δ^{WS}) and mean between-subject differences (δ^{BS}). The differences are expressed as a percentage and bracketing indicates the standard deviation.



(a) (b) (c)

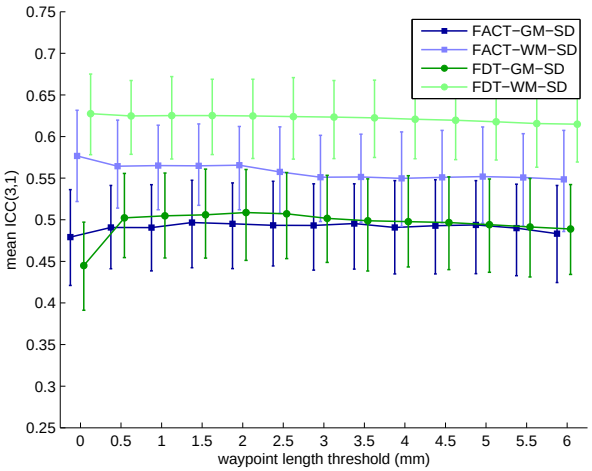
Figure 1:



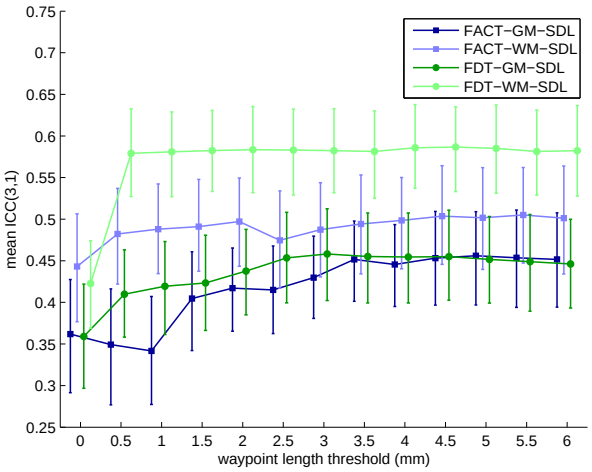
(a) SD-weighted (b) SDL-weighted (c) FA-weighted

(d) SD-weighted (e) SDL-weighted (f) FA-weighted

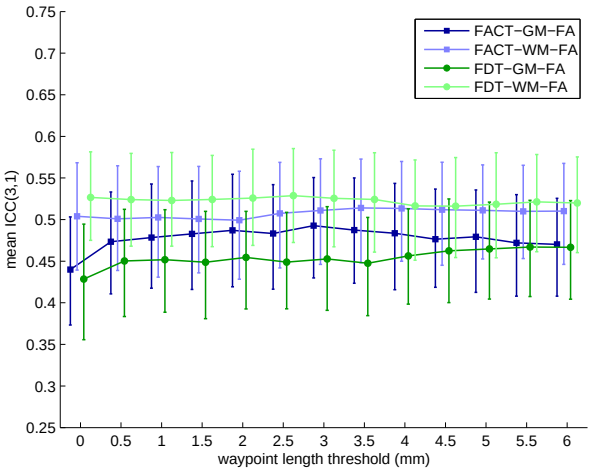
Figure 2:



(a) SD-weighted

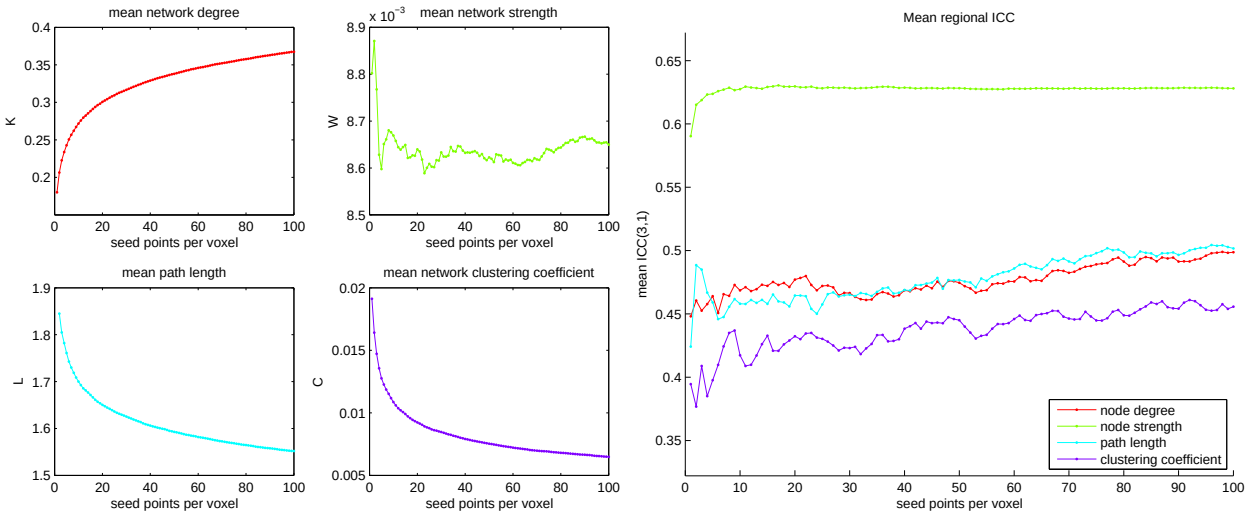


(b) SDL-weighted



(c) FA-weighted

Figure 3:


mean network degree
K
seed points per voxel
x 10^-3
mean network strength
W
seed points per voxel
mean path length
L
seed points per voxel
mean network clustering coefficient
C
seed points per voxel
Mean regional ICC
mean ICC(3,1)
seed points per voxel
node degree
node strength
path length
clustering coefficient

(a) (b)

Figure 4:

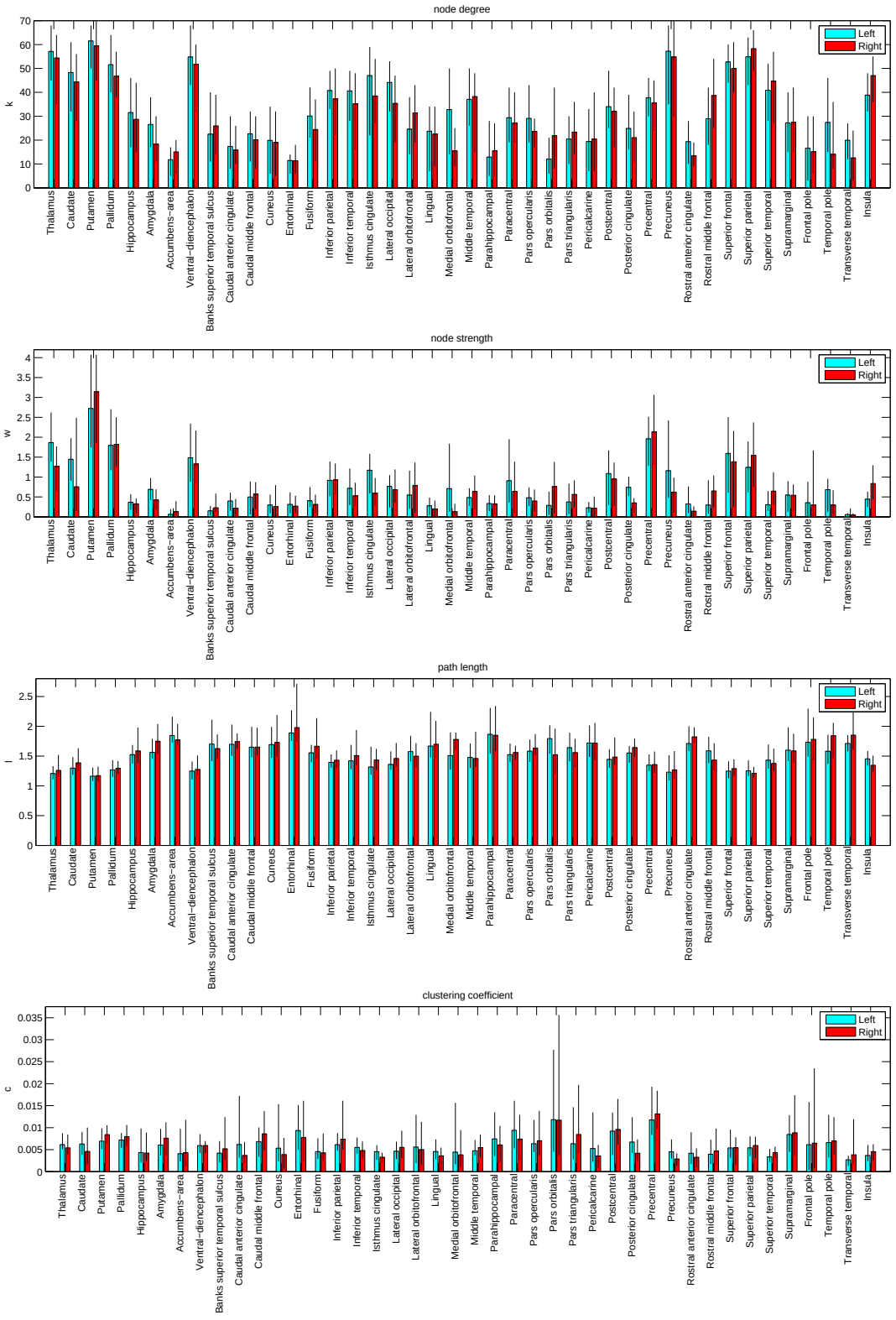

node degree
k
Left
Right
node strength
w
path length
l
clustering coefficient
c
Thalamus
Caudate
Putamen
Pallidum
Hippocampus
Amygdala
Accumbens−area
Ventral−diencephalon
Banks superior temporal sulcus
Caudal anterior cingulate
Caudal middle frontal
Cuneus
Entorhinal
Fusiform
Inferior parietal
Inferior temporal
Isthmus cingulate
Lateral occipital
Lateral orbitofrontal
Lingual
Medial orbitofrontal
Middle temporal
Parahippocampal
Paracentral
Pars opercularis
Pars orbitalis
Pars triangularis
Pericalcarine
Postcentral
Posterior cingulate
Precentral
Precuneus
Rostral anterior cingulate
Rostral middle frontal
Superior frontal
Superior parietal
Superior temporal
Supramarginal
Frontal pole
Temporal pole
Transverse temporal
Insula

Figure 5:

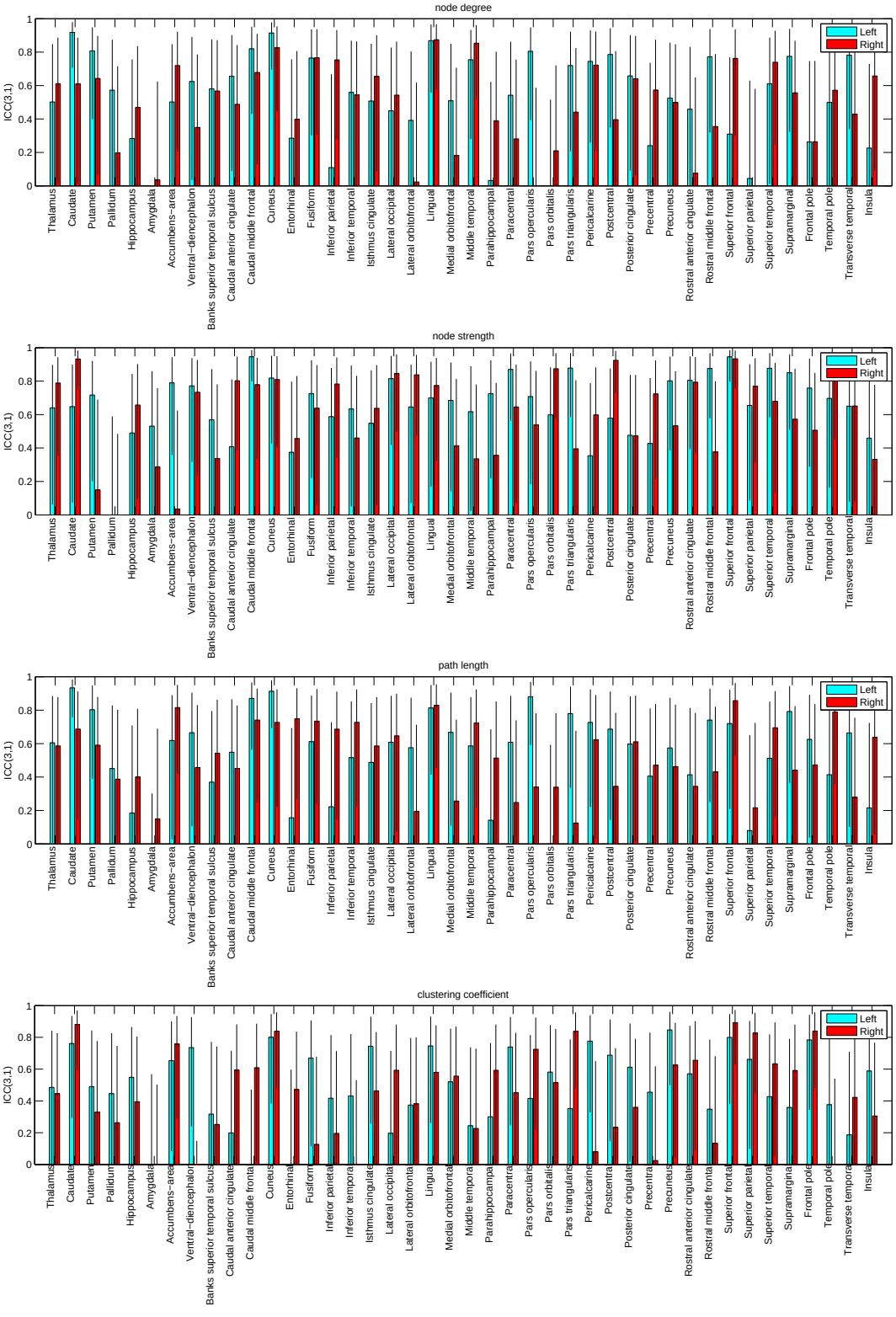

node degree
node strength
path length
clustering coefficient
ICC(3,1)
Left
Right
Thalamus
Caudate
Putamen
Pallidum
Hippocampus
Amygdala
Accumbens−area
Ventral−diencephalon
Banks superior temporal sulcus
Caudal anterior cingulate
Caudal middle frontal
Cuneus
Entorhinal
Fusiform
Inferior parietal
Inferior temporal
Isthmus cingulate
Lateral occipital
Lateral orbitofrontal
Lingual
Medial orbitofrontal
Middle temporal
Parahippocampal
Paracentral
Pars opercularis
Pars orbitalis
Pars triangularis
Pericalcarine
Postcentral
Posterior cingulate
Precentral
Precuneus
Rostral anterior cingulate
Rostral middle frontal
Superior frontal
Superior parietal
Superior temporal
Supramarginal
Frontal pole
Temporal pole
Transverse temporal
Insula

Figure 6:

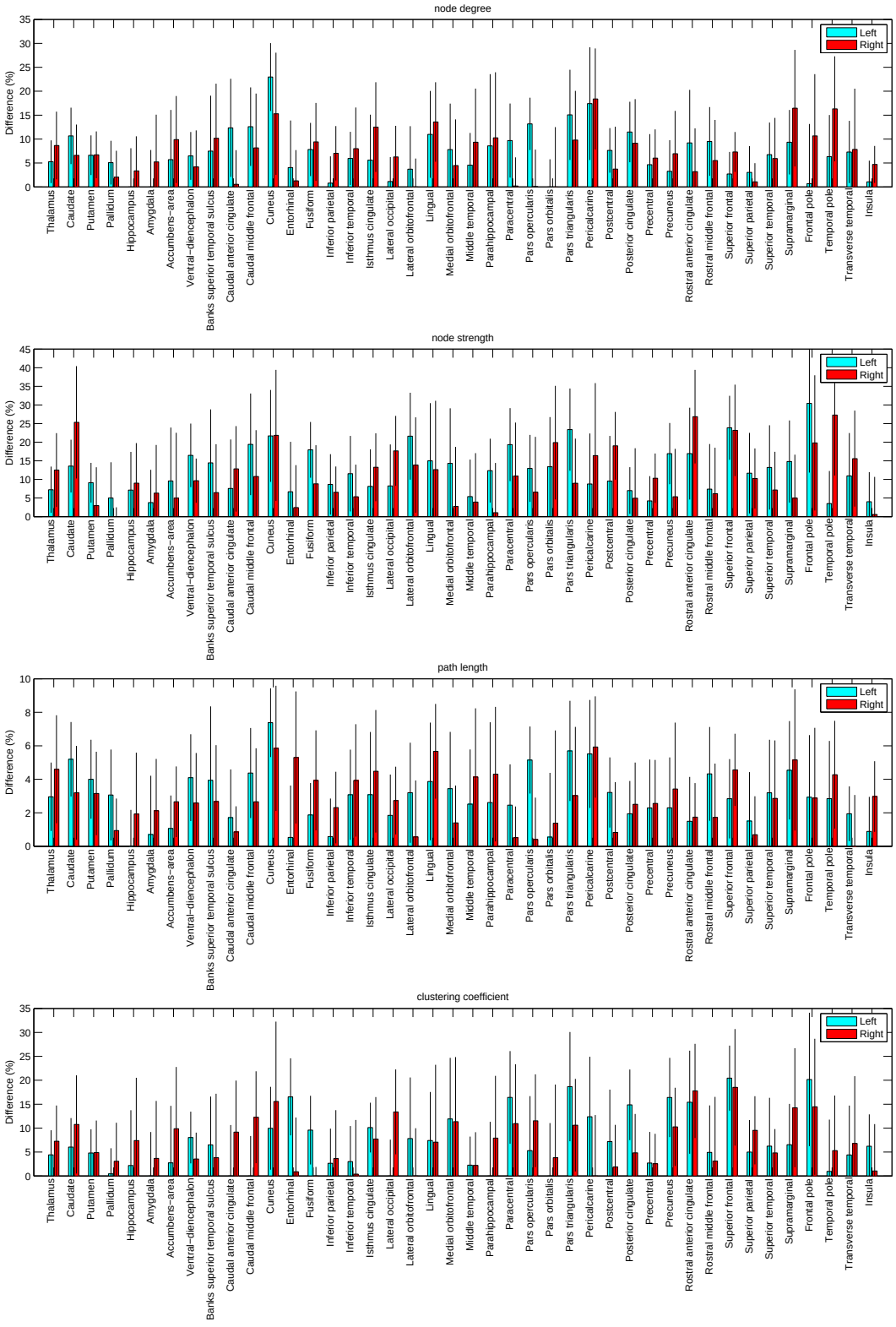

node degree
node strength
path length
clustering coefficient
Difference (%)
Left
Right
Thalamus
Caudate
Putamen
Pallidum
Hippocampus
Amygdala
Accumbens−area
Ventral−diencephalon
Banks superior temporal sulcus
Caudal anterior cingulate
Caudal middle frontal
Cuneus
Entorhinal
Fusiform
Inferior parietal
Inferior temporal
Isthmus cingulate
Lateral occipital
Lateral orbitofrontal
Lingual
Medial orbitofrontal
Middle temporal
Parahippocampal
Paracentral
Pars opercularis
Pars orbitalis
Pars triangularis
Pericalcarine
Postcentral
Posterior cingulate
Precentral
Precuneus
Rostral anterior cingulate
Rostral middle frontal
Superior frontal
Superior parietal
Superior temporal
Supramarginal
Frontal pole
Temporal pole
Transverse temporal
Insula

Figure 7: